Ferdowsi University of Mashhad

**Journal of Systems Thinking in Practice**
**JSTINP, *QUARTERLY***
**Homepage: jstinp.um.ac.ir**
**Online ISSN: 2821-1669, Print ISSN: 2980-9460**
**Research Article**
DOI: 10.22067/jstinp.2026.98779.1206
JSTINP 2026; Vol. 5, No. 3

# A Stakeholder-aware Decision Support System for Make-to-Order Production Planning Using Retrieval-Augmented Generative Artificial Intelligence

Mozhdeh Rahjoo [a]*, Farzad Haghighi Rad [a], Amir-Reza Abtahi [a]

*a Department of Operations Management and Information Technology, Faculty of Management, University of Kharazmi, Tehran, Iran.*

***How to cite this article***

Rahjoo, M., Haghighi Rad, F. and Abtahi, A.-R., 2026. A Stakeholder-aware Decision Support System for Make-to-Order Production Planning Using Retrieval-Augmented Generative Artificial Intelligence. *Journal of Systems Thinking in Practice. 5(3)*, pp.1-25. Doi: 10.22067/jstinp.2026.98779.1206.
URL: https://jstinp.um.ac.ir/article_48516.html

**A B S T R A C T**

Make-to-order (MTO) manufacturing involves production planning decisions under uncertain orders, incomplete operational information, and conflicting stakeholder priorities. To address the need for traceable and defensible cross-functional alignment, this study develops a stakeholder-aware, retrieval-augmented-generation-enabled decision support system (DSS) operationalized through a mixed-methods industrial case study. The system integrates empirical data from 200 orders, 279 downtime events, and 400 quality control, nonconformity, and optical testing records. Rather than relying on probabilistic models for execution calculations, the architecture enforces deterministic release gates to evaluate organisational release readiness prior to optimisation. Retrospective evaluation reveals that the gates classified organisational release status, releasing 170 orders and placing 30 orders on HOLD due strictly to quality control (QC) and optical time-domain reflectometer (OTDR) violations. For the organisationally releasable orders, the engine evaluates two benchmark policies (BASE_FIFO and BASE_EDD) together with three generated scenarios (PROFIT, EFFICIENCY, and CONSENSUS) using the implemented E2 stakeholder-weighted evaluation model. The objective weights were obtained through direct preference elicitation (production efficiency = 0.4688, profit = 0.2997, and value added = 0.2314). The baseline policy (BASE_EDD) benchmark achieved the highest E2 stakeholder-weighted score of 0.5664 and maintained an OTD rate of 90.0% for the evaluated dataset. By restricting the local Large Language Model (Llama-3-8B) strictly to an explainable AI (XAI) interaction layer, the system generates source-grounded justifications without altering deterministic data. The primary academic contribution of this research is the structural separation of organisational release control from multi-objective scenario evaluation, providing a verifiable socio-technical framework for complex make-to-order planning.

***Keywords***

Make-to-order manufacturing, Systems thinking, Decision support system, Retrieval augmented generation, Generative artificial intelligence, Stakeholder Alignment.

***Article history***

Received: 2026-05-17
Revised: 2026-07-13
Accepted: 2026-07-21
Published (Online): 2026-09-28

**Number of Figures: 1** **Number of Tables: 4** **Number of Pages: 25** **Number of References:20**

**Corresponding author*
**Email:** Mozhdehrahjoo@khu.ac.ir

## 1. Introduction

Make-to-order (MTO) manufacturing environments are characterised by high product variety, uncertain order arrivals, short response times, and frequent changes in customer requirements. In such environments, production planning is not only a technical scheduling task; it is a systemic decision-making problem in which production, sales, finance, quality control, and management actors must coordinate their priorities under uncertainty. MTO and engineer-to-order (ETO) planning require cross-functional coordination because delivery commitments depend on operational information and decisions distributed across sales, engineering, procurement, and production functions (Bhalla et al., 2023). Viewed through the lens of Systems Thinking, this multi stakeholder environment requires a sociotechnical approach to resolve inherent interdepartmental conflicts, such as the tension between sales driven delivery agility and production driven efficiency.

This multi stakeholder nature makes MTO planning complex because decisions are shaped by due dates, capacity, material availability, quality readiness, payment conditions, unresolved nonconformities, and operational disruptions. Traditional planning approaches such as material requirements planning (MRP), enterprise resource planning (ERP), and capacity requirements planning (CRP) provide useful foundations, but they often assume relatively stable demand, complete operational data, and predefined routings. Operations research and Digital Twin approaches provide rigorous mechanisms for order acceptance, scheduling, rescheduling, and job allocation under operational constraints. However, these approaches primarily rely on structured numerical inputs and predefined objectives, and generally provide limited support for retrieving unstructured organisational evidence or generating source-grounded managerial explanations (An et al., 2023; Sit and Lee, 2023).

Generative AI and Large Language Models (LLMs) create opportunities for interpreting heterogeneous information, generating planning alternatives, and explaining decision options. However, the direct use of language models in operational planning is risky because outputs may be plausible but unsupported by organisational evidence or inconsistent with execution constraints, a phenomenon known as AI hallucination. Retrieval Augmented Generation (RAG) can support evidence-grounded interaction by retrieving domain-specific and real-time information. Existing manufacturing applications, however, have primarily focused on equipment-level monitoring or general generative artificial intelligence (GenAI) capabilities; the integration of retrieved evidence with deterministic release controls and stakeholder-weighted tactical planning remains underexplored (Jackson et al., 2024; Jeon et al., 2025).

A central design challenge is to prevent a probabilistic language model from performing deterministic calculations, changing hard operational constraints, or introducing unsupported production information. The architecture must therefore separate analytical decision logic from natural-language generation and restrict the generative layer to evidence grounded explanation. Despite growing interest in AI-enabled decision support, an integration gap remains. Existing research has separately examined cross-functional planning, optimisation-based scheduling, Digital Twin-enabled planning, and GenAI-assisted manufacturing. However, limited empirical evidence exists on their integration into a single MTO workflow that combines organisational evidence retrieval, deterministic readiness controls, stakeholder-weighted scenario evaluation, traceable explanation, and final human authorization (Bhalla et al., 2023; Jackson et al., 2024; Leng et al., 2026).

To address this gap, we have explicitly anchored this study within Sociotechnical Systems (STS) Theory and Systems thinking in Practice. This research develops a stakeholder-aware decision support system for MTO production planning, framing our RAG enabled DSS not merely as a computational engine, but as an organisational artefact that operationalizes systemic alignment, decision traceability, and multi-actor boundary-spanning coordination. The specific design contributions of this study are threefold:

1) A layered separation of organisational release conditions from scenario-level performance evaluation through explicit payment, quality-control, OTDR, and NCR controls.
2) A stakeholder-weighted evaluation mechanism that consolidates the elicited E2 objective preferences and supports transparent ranking of the evaluated planning alternatives.
3) A decoupled evidence-grounded architecture in which Generative AI is restricted to natural-language explanation and interaction without authority to modify release decisions, calculated key performance indicators (KPIs), or rule identifiers.

## 2. Literature Review

This section reviews three focused streams of literature—MTO production planning as a multi-stakeholder system, soft-systems-oriented planning and evidence-grounded decision support, and the stakeholder-aware planning research gap.

### *2.1. MTO Production Planning as a Multi-stakeholder System*

MTO manufacturing differs from make-to-stock production because production is triggered by confirmed customer orders rather than forecasted demand. Product specifications, material needs, routings, and delivery requirements may remain uncertain until order acceptance, creating a planning context characterised by high variety, uncertainty, and limited operational visibility (Hörbe and Erol, 2024; Kapulin and Russkikh, 2020).

The complexity of MTO planning is not limited to scheduling. MTO planning involves simultaneous delivery, capacity, and economic considerations (Ma et al., 2024; Mundt and Lödding, 2025). In the present case, the stakeholder interviews further showed that sales, production, finance, and quality-control actors interpreted these considerations through different functional priorities.

### *2.2. Soft Systems thinking for Semi-structured Planning Problems*

Production planning in MTO manufacturing can be viewed as a semi-structured and multi-stakeholder problem because it combines technical constraints with human judgment, organisational priorities, and conflicting interpretations of feasibility and performance. Such problems are not fully addressed through optimisation alone, because the definition of an acceptable plan depends on how different actors interpret delivery reliability, production efficiency, quality readiness, financial conditions, and operational risk. Soft-systems-oriented approaches are relevant in such contexts because they help structure complex problem situations by considering local context, multiple stakeholder perspectives, and value-based objectives rather than assuming a single predefined problem formulation (Augustsson et al., 2020; Françozo et al., 2022).

In this study, the soft systems orientation is used as a conceptual lens for structuring the planning situation rather than as a full implementation of Soft Systems Methodology. The purpose is to support the development of a decision support system that makes stakeholder priorities, operational constraints, evidence sources, and trade-offs explicit. This interpretation is consistent with recent applications of soft systems thinking that emphasise shared understanding, participatory sensemaking, and systemically desirable actions in complex organisational settings (Goto and Miura, 2022; Wattuhewa and Pheerathan, 2025).

A soft-systems-oriented view does not replace analytical planning tools; rather, it helps frame the planning problem as a sociotechnical decision process. In this view, production planning requires mechanisms for structuring stakeholder concerns, making assumptions explicit, comparing alternative interpretations of the problem, and supporting purposeful action. Consequently, in the context of this study, soft systems thinking provides a suitable lens for understanding why MTO planning requires a stakeholder-aware decision support system. The proposed system is not intended to automate planning decisions completely. Instead, it aims to structure a complex planning situation by connecting operational evidence, control rules, scenario evaluation, and stakeholder preferences. This framing is crucial for a systems-oriented journal because it positions the contribution as a systemic decision support mechanism rather than merely an AI-based planning tool.

Building on this conceptual lens, the soft systems orientation was used not merely as a theoretical justification but as a design perspective that informed the architectural inclusion of human-centred review points and release control logic. Rather than treating production planning as a closed deterministic optimisation problem, this perspective supported the explicit representation of multiple stakeholder concerns, organisational evidence, and human authorisation. It also helped incorporate information such as pending payment clearance and unresolved non-conformities, which may not be directly represented in schedulers that rely exclusively on structured numerical inputs. Accordingly, the evaluation extends beyond machine utilisation maximisation and considers organisational readiness, traceability, and stakeholder-weighted performance.

### *2.3. Evidence-grounded Decision Support Using Generative AI and RAG*

Recent developments in Generative AI and Large Language Models (LLMs) create new opportunities for decision support in manufacturing and operations management. These technologies can interpret heterogeneous information, summarise organisational knowledge, and support natural-language interaction. In decision-support settings, they may also assist planners in formulating or explaining alternative courses of action, provided that the underlying calculations are delegated to appropriate analytical tools (Garcia et al., 2024; Jackson et al., 2024).

However, the direct use of Large Language Models in operational decision-making is risky. Generated outputs may be linguistically plausible but unsupported by valid evidence or inconsistent with shop floor constraints. In production planning, this risk is critical because a recommendation may appear reasonable while still violating material availability, quality readiness, payment conditions, unresolved non-conformities, or capacity limitations. These limitations motivate the use of controlled architectures in which generative models are combined with domain data, governance mechanisms, and external analytical tools rather than being assigned unrestricted numerical decision authority (Ghobakhloo et al., 2024; Kusiak, 2025).

Retrieval-Augmented Generation (RAG) can ground generated outputs in retrieved organisational evidence. In industrial settings, this capability is particularly relevant because decisions often depend on company-specific data, rules, procedures, historical records, and contextual knowledge that may not be represented in a general-purpose language model. Tang *et al.* (2024) demonstrate that RAG can introduce context-specific constraints into a scheduling workflow, while Jeon *et al.* (2025) show how real-time industrial data can ground conversational machine monitoring. Building on these capabilities, the present study introduces an additional deterministic release-control layer through which retrieved evidence is checked against company-specific financial and quality rules before an order enters the scheduling stage.

Human review remains important in this type of decision-support system, but it should be defined carefully. Human-centred XAI research emphasises that explanations should support appropriate understanding, trust calibration, and responsible interaction rather than merely expose algorithmic details (Cortiñas-Lorenzo et al., 2025; Nikiforidis et al., 2025). Accordingly, the proposed artefact includes explicit review points at which planners or relevant stakeholders examine cases involving incomplete evidence, conflicting control conditions, or competing scenario outcomes.

### *2.4. Stakeholder-Aware Planning and Research Gap*

Scenario-based evaluation can support decision-makers in comparing the consequences of alternative planning logics rather than relying on a single recommended output. Related research has emphasised the value of GenAI-assisted decision support and data-driven production-planning alternatives, although stakeholder-weighted comparison remains less developed (Jackson et al., 2024; Ma et al., 2024).

The reviewed literature indicates that optimisation-based planning, cross-functional coordination, manufacturing LLMs, and RAG-enabled industrial assistance have largely been examined as complementary but separate research streams. Table 1 presents a structured comparison of representative related approaches and clarifies the integration boundary addressed by the proposed artefact.

Table 1. Comparison of representative related approaches and the proposed artefact

| Study | Decision context and core approach | Cross-functional or stakeholder scope | Evidence grounding and release controls | Main contribution and remaining boundary |
|---|---|---|---|---|
| An et al. (2023) | Real-time order acceptance and flexible job-shop rescheduling through integrated mathematical optimisation with maintenance constraints. | Stakeholder perspectives are not explicitly represented; decisions are modelled through predefined objectives and operational constraints. | No RAG-based retrieval of organisational evidence. Operational constraints are embedded in the optimisation model, but documentary financial and quality release rules are not examined. | Provides rigorous joint optimisation of order acceptance and rescheduling; does not address unstructured organisational evidence or source-grounded managerial explanation. |
| Sit and Lee (2023) | Low-volume, high-mix job allocation and production scheduling through a Digital-twin-enabled approach. | Primarily concerned with operational resources, job allocation, and production execution. | Grounded in Digital twin and production data rather than document-based RAG. Explicit financial or documentary quality-release gates are not implemented. | Demonstrates Digital twin support for job allocation and scheduling; cross-functional evidence governance and stakeholder-weighted explanation remain outside the evaluated scope. |
| Bhalla et al. (2023) | Delivery-date setting and sales and operations planning (S&OP) in engineer-to-order manufacturing through a systematic literature synthesis and reference framework. | Explicitly incorporates sales, engineering, procurement, and production perspectives. | Uses cross-functional planning information, but does not implement a RAG evidence layer or executable release gates. | Establishes the need for coordinated cross-functional planning; does not develop or evaluate a GenAI-enabled planning artefact. |
| Jackson et al. (2024) | Generative AI in supply chain and operations management through a capability-based conceptual framework. | Considers organisational capabilities and implementation implications. | Discusses information-related and GenAI capabilities, but does not implement a case-specific RAG pipeline or organisational release-control mechanism. | Provides a broad framework for analysing GenAI capabilities; does not evaluate deterministic release controls or stakeholder-weighted MTO planning. |
| Garcia et al. (2024) | LLM-based manufacturing assistance through a manufacturing LLM framework and document-grounded assistant. | Primarily supports user-system interaction and manufacturing-information access. | Uses uploaded documents and context-specific manufacturing information. External analytical agents are recommended for mathematical tasks, but organisational release gates are not implemented. | Demonstrates document-grounded manufacturing assistance and identifies limitations of LLM-only calculation; tactical planning and stakeholder-weighted scenario ranking are not evaluated. |
| Tang et al. (2024) | Conversion of static scheduling models into dynamic scheduling workflows through | Focuses on user-defined scheduling changes rather than cross- | Retrieves problem-specific scheduling knowledge through RAG. Scheduling constraints are | Shows how RAG can support context-specific scheduling modifications; organisational evidence governance and |

| | | | | |
|---|---|---|---|---|
| | RAG-assisted generation of constraints from natural language. | functional stakeholder alignment. | incorporated, but financial and quality-document release controls are not examined. | multidisciplinary release authorisation remain outside its scope. |
| **Jeon et al. (2025)** | Real-time CNC machine monitoring through a multi-agent LLM framework with real-time RAG. | Focuses on human-machine interaction at the equipment level. | Retrieves real-time IIoT and CNC-machine data. Cross-functional payment, QC, OTDR, and NCR release gates are not implemented. | Demonstrates reliable equipment-level conversational monitoring; cross-functional tactical planning and stakeholder-weighted scenario comparison are outside its evaluated scope. |
| **Proposed artefact** | MTO production release and scenario-based planning through deterministic release controls, rule-based scenarios, stakeholder-weighted evaluation, card-based evidence retrieval, and GenAI explanation. | Integrates production, planning, sales, finance, quality control, and management perspectives. | Retrieves traceable rule and fact cards from an organisational knowledge repository and evaluates payment, QC, OTDR, and NCR controls before scheduling; in the reported execution, the observed HOLD decisions were triggered by QC/OTDR conditions. | Implements and retrospectively evaluates an integrated workflow that separates organisational release control, scenario evaluation, evidence retrieval, generative explanation, and final human authorisation. |

As shown in Table 1, prior studies provide important but complementary capabilities. Optimisation and Digital twin approaches offer rigorous support for order acceptance, allocation, rescheduling, and production execution; cross-functional planning research identifies the information flows required across organisational functions; and manufacturing LLM and RAG studies demonstrate document-grounded assistance, dynamic constraint generation, and real-time equipment interaction. However, limited empirical evidence exists on integrating these capabilities into a single MTO workflow that separates organisational release control from scenario-level performance evaluation and restricts the language model to evidence-grounded explanation. The proposed artefact addresses this integration gap by combining explicit financial and quality release controls, stakeholder-weighted comparison of planning alternatives, traceable organisational evidence, and final human authorisation.

## 3. Method

This study adopts a systems thinking and sociotechnical approach to develop and evaluate a decision support system for production planning in a MTO manufacturing context. Rather than treating the research problem as a purely technical scheduling issue, it is framed as a semi-structured, multi-stakeholder decision situation. This orientation is necessary because MTO planning involves conflicting stakeholder objectives, incomplete operational information, and

 DOI: 10.22067/jstinp.2026.98779.1206

strict organisational constraints that cannot be resolved through automated mathematical optimisation alone.

### *3.1. Research Approach and Data Sources*

This study uses a mixed-methods industrial case-study design at a make-to-order fibre-optic cable manufacturer (Tida Ertebat Niayesh). The term “mixed methods” refers strictly to the strategic integration of quantitative operational records and qualitative stakeholder interviews to inform the development and retrospective evaluation of the system. Specifically, quantitative data govern the feasibility-screening gates and numerical scenario evaluations, while qualitative data capture heuristic planning rules and expert priority weights. Design science research (DSR) is not claimed as a formal methodological protocol; instead, the research strategy is bounded entirely within an empirical, case-based sociotechnical evaluation framework.

The system architecture and logic were built and evaluated using empirical data spanning an 18-month operational period. The quantitative dataset includes 200 order records, 279 production-downtime events and 400 records relating to quality control (QC), non-conformity reports (NCRs) and optical time-domain reflectometry (OTDR). The qualitative data were collected through six semi-structured interviews with functional heads in production, sales, finance, quality control and management.

### *3.2. Architecture of the Decision Support System*

As illustrated in Figure 1, the conceptual architecture of the proposed stakeholder-aware decision support system is designed to address the limitations of traditional planning models. Instead of relying on generic frameworks, we developed a highly traceable, four-stage decision pipeline was developed to reflect the sociotechnical realities of the factory. These architectural stages, which systematically transform unstructured organisational rules into executable plans, are summarised in Table 2 and detailed as follow:

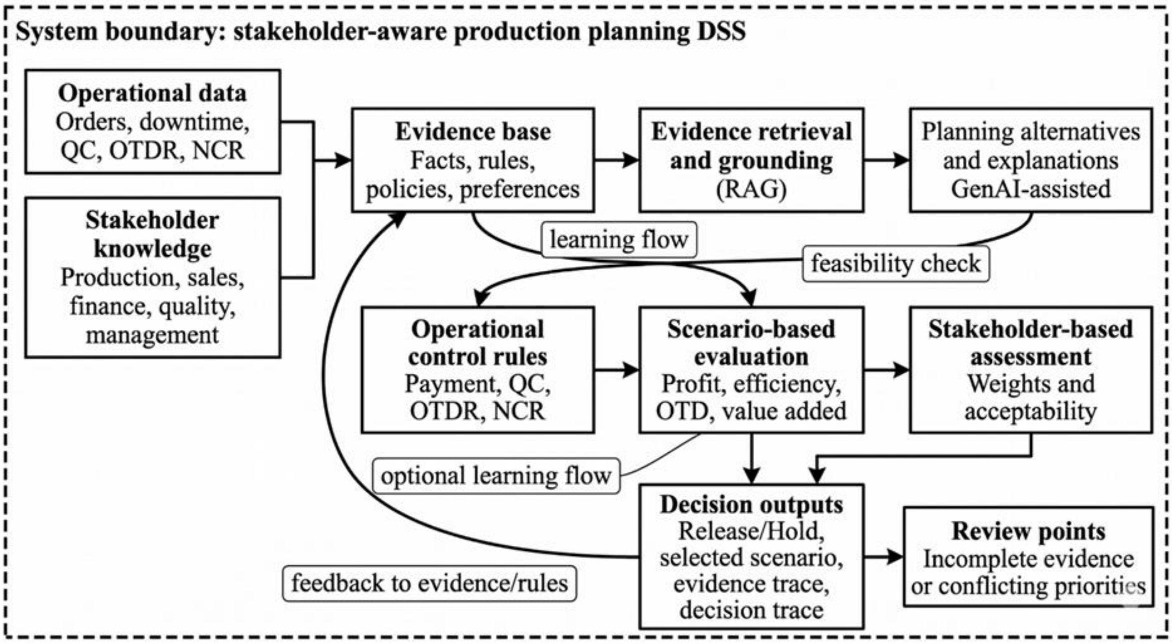


Figure 1 .Conceptual architecture of the proposed stakeholder-aware decision support system.

Table 2. The four-stage architecture of the stakeholder-aware DSS

| Architectural Stage | Systemic Function | Key Components and Logic | Operational Output |
|---|---|---|---|
| 1. Constraint Isolation (Release Gates) | Separating hard operational constraints from performance KPIs. | Validation against Payment, QC, OTDR, and NCR status. | Feasible valid orders; held orders with justification codes. |
| 2. Scenario and KPI Evaluation | Evaluating production alternatives against operational goals. | Scenario rules (FIFO, EDD, Profit-Max); KPI calculations. | Multiple generated schedules and multi-objective KPI metrics. |
| 3. Stakeholder-Weighted E2 Evaluation | Making stakeholder-derived objective priorities explicit in the comparison of candidate planning options. | Stakeholder-derived E2 objective weights; implemented E2 evaluation model; comparison and ranking of candidate options. | Ranked candidate options submitted for human review and authorisation. |
| 4. Evidence Retrieval and HITL | Grounding mathematical outputs in verifiable organisational reality. | RAG Knowledge Cards; Explainable AI (XAI) translation layer. | Traceable structured JSON; Natural-language explanation. |

*Stage 1: Constraint Isolation via Release Gates*

Unlike traditional models that treat quality or payment issues as post-generation KPIs, our system enforces 'Release Gates' prior to planning. Incoming orders are validated against hard constraints (Payment status, QC approval, OTDR results, and open NCRs). Orders failing any gate are immediately flagged as “HOLD “and assigned a specific reason code, preventing orders blocked by predefined financial or quality conditions from entering the scheduling stage.

*Stage 2: Scenario Generation and KPI Evaluation*

For orders that pass the Release Gates, the engine generates three planning scenarios (PROFIT, EFFICIENCY, and CONSENSUS) and evaluates them together with two reproducible benchmark policies, the earliest-due-date policy (BASE_EDD) and the first-in, first-out policy (BASE_FIFO).

*Stage 3: Stakeholder-Weighted Evaluation*

The relative importance of production efficiency, profit, and value added was elicited through six semi-structured interviews with functional representatives. The elicited judgements were consolidated and reviewed to obtain the final E2 objective weights: production efficiency = 0.4688, profit = 0.2997, and value added = 0.2314. These weights were applied through the implemented E2 evaluation model to generate a common stakeholder-weighted score for each candidate option. The resulting composite scores were obtained directly from the implemented case-study calculation model and were retained unchanged from the validated research outputs. The scores were used to support transparent comparison of the candidate options and did not replace final human authorisation.

*Stage 4: Evidence Retrieval and Human-in-the-Loop Review (HITL)*

The system grounds its output using RAG-style knowledge cards. Every application of an operational rule is associated with a rule identifier, source and categorical verification status. This structure supports decision traceability while preserving final authorisation for the human decision-maker.

### *3.3. Algorithmic Logic and Pseudocode*

Both quantitative and qualitative data were used. Quantitative data included order records, production-downtime events and quality-control records. Qualitative data were collected through semi-structured interviews with stakeholders involved in planning and decision-making. The quantitative data supported feasibility checks and scenario evaluation, while the qualitative data captured planning rules, stakeholder priorities and decision requirements that were not fully represented in operational databases.

To provide a structured overview of the procedural methodology, the principal processing logic of the decision engine is summarised in Algorithm 1. This algorithm demonstrates how unstructured organisational rules are transformed into executable code, governing the flow from Release Gates validation to final scenario selection.

**Algorithm 1: Stakeholder-aware MTO Production Planning with Release Gates**

Inputs:

Order List (O);

Operational Data (Payment, QC, OTDR, NCR);

Stakeholder-Derived Objective Weights (Weff, Wprof, Wva)

Outputs:

Highest-Ranked Candidate Option (S*);

E2 Stakeholder-Weighted Score;

Evidence Log

```
Initialize Benchmark_Policies = [BASE_FIFO, BASE_EDD]
Initialize Generated_Scenarios = [PROFIT, EFFICIENCY, CONSENSUS]
Initialize Candidate_Options = Benchmark_Policies + Generated_Scenarios
Initialize Hold_List = [], Valid_Orders = [], Evidence_Log = []
// Stage 1: Release-Gate Validation
for each order o in O do
    if Payment(o) == PENDING or QC(o) == FAIL
       or OTDR(o) == FAIL or NCR(o) == OPEN then
        Append o to Hold_List
        Append Retrieve_Knowledge_Card(o, Reason) to Evidence_Log
    else
        Append o to Valid_Orders
    end if
end for
// Stage 2: Candidate-Option Generation and Raw KPI Computation
for each option c in Candidate_Options do
    Schedule(c) = Generate_Schedule(Valid_Orders, rule = c)
    Raw_KPI_Profit(c) = Calculate_Total_Profit(Schedule(c))
```

```
    Raw_KPI_Eff(c) = Calculate_Efficiency(Schedule(c))
    Raw_KPI_VA(c) = Calculate_Value_Added(Schedule(c))
end for
// Stage 3: Stakeholder-Weighted E2 Evaluation
for each option c in Candidate_Options do
    E2_Score(c) =
      Apply_Implemented_E2_Evaluation (
        Raw_KPI_Profit(c),
        Raw_KPI_Eff(c),
        Raw_KPI_VA(c),
        Weff,
        Wprof,
        Wva )
end for
// Stage 4: Ranking and Evidence Compilation
S* = argmax(E2_Score)
Evidence_Log =
    Compile_Knowledge_Cards(S*, Hold_List, Evidence_Log)
Return S*, E2_Score(S*), Evidence_Log
```

The implemented E2 evaluation model combines the scenario-level objective indicators with the stakeholder-derived objective weights to provide a common comparison score for the candidate options. The revised manuscript reports the raw scenario-level KPIs, the final objective weights, and the composite E2 scores produced by the implemented case-study decision engine. Because the complete transformation parameters were embedded in the original calculation model, the study does not characterise the procedure as analytic hierarchy process (AHP) or as option-wise Min–Max normalisation.

### *3.4. Technical Implementation and the Role of Generative AI*

To ensure reproducibility and operational reliability, the technical framework of the proposed DSS employs a decoupled architecture, intentionally separating deterministic calculations from generative explanations.

The Deterministic Core: The back-end decision engine, release-gate logic and KPI computations were developed entirely in Python (version 3.11) using the pandas library. The back end is exposed through a representational state transfer (REST) application programming interface (API) using the FastAPI framework, with endpoints such as /plan and /release-gates. This separation ensures that schedules and consensus scores are not generated by the LLM and are therefore not subject to generative-model variability.

The RAG-Enabled Generative AI Layer: In line with the "RAG-Enabled Generative AI Approach", the system uses a Llama-3-8B-Instruct model, executed locally through Ollama, strictly as an explainable artificial intelligence (XAI) and natural-language interaction layer. Once the deterministic core identifies the highest-ranked candidate option under the stakeholder-derived weights and retrieves the relevant knowledge cards, the structured JavaScript Object Notation (JSON) payload is passed to the LLM. The model's sole responsibility is to generate a human-readable, evidence-grounded explanation for the human decision-maker while explicitly citing the retrieved rules. This architecture is designed to give the planner traceable insights without overriding hard operational constraints.

Hardware Specifications: The computational evaluation of the 200 historical orders, tabular scenario generation and local execution of the Llama-3-8B-Instruct interaction layer were performed on a workstation equipped with an NVIDIA RTX 4090 graphics processing unit (GPU). This configuration demonstrates the technical feasibility of local execution for the evaluated case and avoids transferring the case-study data to an external cloud service. Broader scalability must be evaluated using larger datasets and alternative hardware configurations.

#### *Algorithmic Classification and Execution Metrics*

To address the specific procedural taxonomy requested, the developed system does not rely on stochastic metaheuristics or intensive mathematical solvers. Instead, the computational logic is strictly categorised across the following decoupled layers:

1) Gate Screening: Implemented as a deterministic rule-based screening mechanism.
2) Scenario Generation: Executed via rule-based/heuristic sequencing using standardised dispatching rules and benchmark policies (specifically BASE_FIFO and BASE_EDD).
3) Stakeholder Ranking: Operationalised through the implemented E2 evaluation model using the stakeholder-derived objective weights to compare and rank the candidate planning options.
4) Generative Model**:** The local LLM acts exclusively as an explainable AI (XAI) interaction layer for natural-language generation and is strictly prohibited from participating in scheduling calculations or modifying operational metrics.

To evaluate the operational efficiency of this decoupled architecture, computational execution times were systematically audited over 50 independent end-to-end runs using an 18-month historical dataset on a local workstation with an Intel Core i9-13900K central processing unit (CPU), 64 GB of double data rate 5 (DDR5) random-access memory (RAM) and a single NVIDIA RTX 4090 GPU with 24 GB of video random-access memory (VRAM). To eliminate transient latency anomalies, a five-iteration warm-up phase was discarded before log recording. The local generative layer deployed the Llama-3-8B-Instruct model in unquantised 16-bit floating-point format (FP16), using approximately 16 GB of VRAM to optimise inference speed within a maximum context length of 4,096 tokens.

Under this controlled evaluation profile, the deterministic core processed the 200-order batch in a mean execution time of 0.12 seconds (an average of 0.6 milliseconds per order for joint hard-gate screening, scenario generation and multi-criteria evaluation; standard deviation, $\sigma$ = 0.01 seconds). The local generative-explanation phase required a mean execution time of 8.40 seconds to compile the complete evidence-cited textual summary (standard deviation, $\sigma$ = 0.28 seconds). Consequently, the mean total end-to-end processing time for a full tactical batch was 8.52 seconds, indicating low computational latency for the evaluated retrospective batch; real-time shop-floor readiness requires prospective deployment testing.

To provide a clear and structured overview of the system's technological foundation, Table 3 summarises the complete software and hardware stack used in the implementation of this study.

 DOI: 10.22067/jstinp.2026.98779.1206

Table 3. Technical stack and implementation components

| Component | Technology |
|---|---|
| Programming Language | Python 3.11 |
| Back-end API | FastAPI |
| Data Processing | Pandas |
| Knowledge Base | JSON/Knowledge Cards |
| RAG Component | Evidence Retrieval Layer |
| LLM | Llama-3-8B-Instruct |
| LLM Runtime | Ollama |
| User Interface | Streamlit |
| Hardware | NVIDIA RTX 4090 |

## 4. Results

To address the multi-objective complexity of MTO production and evaluate the technical artefact, the proposed DSS was applied to 200 historical order records alongside 400 operational quality and financial records. The execution results are reported systematically across the architectural stages.

### *4.1. Stage 1: Constraint Isolation Output (Release Gates)*

Of the 200 sampled orders, 170 orders (85%) passed the release controls and 30 orders (15%) were classified as HOLD. In this retrospective execution, all 30 HOLD decisions were associated with failed QC/OTDR conditions and were recorded under the quality-gate reason code. No sampled order failed the payment or NCR controls. Payment and NCR rules remained active in the rule repository but did not trigger a HOLD decision in this execution. This structural constraint isolation prevented orders blocked by predefined organisational release conditions from entering the scenario-generation and evaluation stage.

### *4.2. Stage 2 and 3: Scenario Evaluation and Stakeholder-Weighted Aggregation*

For the 170 organisationally releasable orders, the decision engine evaluated three generated planning scenarios together with the BASE_EDD and BASE_FIFO benchmark policies. The relative importance of the principal planning objectives was obtained through direct preference elicitation during six semi-structured interviews. The elicited judgements were consolidated and reviewed to obtain the final E2 objective weights: production efficiency = 0.4688, profit = 0.2997, and value added = 0.2314. These weights were applied through the implemented E2 evaluation model to calculate the stakeholder-weighted score for each candidate option. Because the weighting procedure involved direct preference elicitation rather than formal pairwise-comparison matrices, no AHP Consistency Ratio was calculated. Table 4 reports the raw performance indicators and final E2 scores generated by the implemented decision engine.

Table 4. Quantitative comparison of benchmark policies and generated planning scenarios

| Candidate option | Total Profit (Billion IRR) | On-Time Delivery (%) | Capacity Efficiency (%) | E2 Stakeholder-Weighted Score |
|---|---|---|---|---|
| BASE_EDD | 524.546 | 90.0% | 82.07% | 0.5664 |
| BASE_FIFO | 524.475 | 89.0% | 81.75% | 0.0000 |
| Profit-oriented | 524.533 | 72.5% | 82.04% | 0.4936 |
| Efficiency-oriented | 524.484 | 78.5% | 81.80% | 0.0842 |
| Consensus-oriented | 524.522 | 74.5% | 82.01% | 0.4304 |

To evaluate the numerical differences reported in Table 4, the relative changes in the generated scenarios were calculated against the baseline policy (BASE_EDD). Across the alternative pathways, total profit changed by −0.003% to −0.012% and capacity efficiency by −0.037% to −0.328%, whereas on-time delivery (OTD) declined by 11.5 to 17.5 percentage points relative to BASE_EDD Regarding the statistical significance of these differences, the reported metrics represent deterministic aggregate outputs calculated from a single, retrospective industrial dataset rather than independent repeated experimental samples. Consequently, the manuscript does not interpret these variances through probabilistic significance tests (e.g., p-values). The results are strictly characterised as descriptive case-study comparisons. Repeated-run sensitivity analysis, multi-period validation, and multi-case statistical testing are identified as critical trajectories for future prospective research.

### *4.3. Stage 4: Traceability and Explainable AI (XAI) Output*

A critical requirement for DSS adoption in manufacturing is the elimination of "black-box" recommendations. Stage 4 of the system outputs a traceable JSON payload containing the selected scenario, computed KPIs and retrieved knowledge cards. To eliminate probabilistic ambiguity and prevent misleading interpretations of numerical metrics, the structured JSON payload uses explicit categorical validation-status strings instead of arbitrary indices. This structural mapping ensures that the explanation layer receives non-probabilistic, accountable parameters fixed a priori. A sample output generated by the core deterministic engine is presented below:

JSON

```
{
 "selected_scenario": "BASE_EDD",
 "e2_stakeholder_weighted_score": 0.5664,
 "kpis": {
   "Efficiency": "82.07%",
   "Total_Profit_Billion_IRR": 524.546,
   "On_Time_Delivery": "90.0%"
  },
 "release_gates": {
   "total_orders": 200,
   "released_orders": 170,
   "held_orders": 30
  },
 "evidence": [
   {"rule_id": "RC-PAY-50-50", "source": "Finance_SOP",
"verification_status": "programmatic_match"},
   {"rule_id": "RC-QC-OTDR-HOLD", "source": "Quality_Manual",
"verification_status": "expert_validated"}
  ]
}
```

Within the system architecture, a status of “programmatic match” denotes a successful automated database match, whereas “expert validated” indicates that the corresponding organisational rule was reviewed and approved by authorised domain experts. This decoupled framework ensures that the text-generation layer operates strictly on transparent, predefined verification categories rather than dynamic model probabilities.

To translate this deterministic output for the human decision-maker, the generative AI interaction layer (Llama-3-8B-Instruct) processes the JSON payload and generates an evidence-grounded, natural-language explanation. In this illustrative example, the generated execution output synthesised the supplied parameters without introducing unsupported operational data:

As your production-planning assistant, I have reviewed the latest outputs from the decision engine. The system has selected the BASE_EDD scenario as the highest-ranked planning option, achieving an E2 stakeholder-weighted score of 0.5664. This option balances our operational goals by delivering an efficiency rate of 82.07%, total profit of 524.546 billion Iranian rials (IRR) and on-time delivery of 90.0%. During release-gate validation, 170 of the 200 orders were released for production, while 30 were placed on hold. The organisational evidence shows that these holds were enforced because quality-manual constraints, such as failed OTDR bounds, prevented the release of unverified orders.

This practical outcome illustrates how integrating industrial expertise with a RAG-enabled generative AI layer can support a more traceable and auditable decision-making process across cross-functional operations.

## 5. Discussion

### *5.1. Explaining the Results through the Principle of Suboptimisation*

The empirical findings illustrate the systems concept of suboptimisation within the evaluated manufacturing context. As presented in Table 4, policies oriented towards individual functional objectives did not improve the overall combination of outcomes for this specific dataset: while the profit-oriented and efficiency-oriented scenarios showed marginal fluctuations in their primary metrics, they were associated with a substantial reduction in on-time delivery (OTD) to 72.5% and 78.5%, respectively, compared to the 90.0% achieved by the baseline EDD policy. This behavior suggests that in a highly constrained MTO environment, optimising a single functional subsystem can introduce unintended trade-offs across interconnected delivery networks. While traditional optimisation models often prioritise a single mathematical objective function, a soft systems lens suggests that multi-departmental actors may reject configurations that compromise critical delivery targets. The E2 stakeholder-weighted score of BASE_EDD (0.5664) indicates that, for this specific operational profile, this benchmark was more closely aligned with the elicited objective priorities than the other evaluated options.

 DOI: 10.22067/jstinp.2026.98779.1206

### *5.2. The Sociotechnical Role of Release Gates*

Furthermore, identifying 30 non-executable orders before planning highlights a shift from descriptive scheduling to proactive decision support. By applying release gates, the system separates organisational release conditions (such as the failed QC/OTDR conditions observed in this execution) from performance indicators used to compare the planning alternatives. Payment and NCR controls remained active but did not trigger HOLD decisions in the evaluated dataset. This separation explains why orders blocked by the observed QC/OTDR conditions were excluded before scenario generation. In this way, the artefact provides a shared and traceable basis for discussing cross-functional release decisions before planning alternatives are evaluated.

Rather than assuming that a quantitative model can automatically eliminate organisational friction, the decoupled architecture is designed to make the underlying operational constraints explicit. As a boundary-spanning artefact, it shifts the decision-making dynamic from speculative arguments (e.g., sales staff pressing for immediate delivery while the quality function withholds non-conforming orders) to an auditable evaluation of verified operational readiness before optimisation begins.

### *5.3. Theoretical and Managerial Implications*

Theoretically, this study contributes to the DSS and management literature by operationalising systems thinking in practice through a layered decision-support architecture. Using RAG-enabled generative AI as an explanation and interaction layer, rather than as an autonomous planner, is designed to reduce the risk of unsupported generation. The language model receives a structured evidence payload and is not authorised to modify deterministic release decisions, calculated KPIs or rule identifiers. This separation supports more transparent and auditable AI-assisted decision processes, consistent with recent human-centred XAI research (Cortiñas-Lorenzo et al., 2025; Nikiforidis et al., 2025).

Managerially, this architecture fundamentally redefines the production planner's role. Instead of manually resolving conflicting functional logics or reacting to shop-floor infeasibilities, the planner becomes a strategic human-in-the-loop evaluator who authorises evidence-grounded scenarios. Organisations adopting AI for production planning must recognise that, without a deterministic core (release gates) and explicit stakeholder-consensus rules, generative AI recommendations remain difficult to justify or execute in complex, multi-stakeholder MTO environments.

## 6. Conclusion

This study developed and retrospectively evaluated a stakeholder-aware Decision Support System (DSS) to address the sociotechnical complexities of (MTO) production planning. Moving beyond traditional optimisation models that treat execution constraints as post-generation penalties, this research introduced a layered architecture anchored in Systems thinking. By operationalizing explicit 'Release Gates', the system retrospectively isolated nonexecutable orders, capturing cases where 15% of historical orders carried unresolved quality non-conformities. Furthermore, the stakeholder-weighted E2 evaluation mechanism provided a structured framework for comparing candidate planning options against the elicited objective priorities. The E2 model incorporated profit, value added, and production efficiency into the stakeholder-weighted comparison, while OTD and tardiness were retained as consequence indicators for interpreting the operational trade-offs associated with each option.

The primary theoretical contribution of this research is bridging the gap between data-driven execution and systemic stakeholder acceptability. This was achieved by demonstrating how RAG and Generative AI can be safely integrated into industrial planning. By restricting the local language model (Llama-3-8B-Instruct) strictly to an evidence-grounded explanation layer, the architecture is designed to reduce the risk of unsupported generative outputs while preserving the deterministic authority of the release-control and KPI-calculation layers.

Managerially, the proposed system alters the production planner's tactical environment, supporting their role as a strategic Human in the Loop (HITL) evaluator rather than a manual schedule builder. By providing a shared, evidence-grounded framework, the DSS aims to clarify cross-functional trade-offs between sales, production, quality, and finance departments, ensuring that final alternatives are evaluated against documented financial, quality, and operational parameters.

### *6.1. Limitations and Future Research*

Despite these substantive contributions, this study acknowledges certain limitations that present compelling avenues for future research. First, the evaluation was conducted retrospectively using historical operational data from a single MTO manufacturing context. Because the reported metrics represent aggregate outputs from a unique industrial dataset rather than independent repeated experimental samples, the findings serve as descriptive case-study comparisons rather than universal statistical proof. Future studies should focus on longitudinal, prospective real-time deployment on the shop floor to measure behavioral impacts, multi-period validation, and systematic sensitivity analysis under dynamic operational conditions.

Second, the stakeholder-derived weights reflect the preferences elicited from six functional representatives in a single industrial case. Although the consolidated values were reviewed for contextual plausibility, no formal pairwise-consistency test was applicable to the direct preference elicitation procedure. Future research should examine weight stability through repeated elicitation, prospective sensitivity analysis, and validation across alternative decision makers and manufacturing settings. Third, the effectiveness of this DSS remains highly contingent upon the initial completeness and accuracy of the underlying organisational databases and documented Standard Operating Procedures (SOPs).

Although the manuscript reports the stakeholder-derived objective weights, scenario-level KPIs, and final E2 scores, the complete transformation parameters remained embedded in the original case-study calculation model. Future research should provide an independently packaged scoring module, intermediate transformed values, and systematic sensitivity analysis to support full external replication.

Finally, the current architecture uses a single generative AI model strictly to explain deterministic outputs. A promising direction for future research is the exploration of multi-agent large language models (LLMs) in sociotechnical systems. Future iterations could deploy distinct, specialised AI agents representing different functional departments (e.g., a "sales agent" negotiating with a "quality agent") to simulate stakeholder-specific arguments and multi-criteria negotiation processes dynamically before presenting candidate options for human authorisation.

**Acknowledgements**

The authors acknowledge the participating organisation and interviewees for their contribution to the case study.

**Disclosure Statement**

No potential conflict of interest was reported by the author(s).

**Abbreviations**

AHP: Analytic hierarchy process

AI: Artificial Intelligence

API: Application Programming Interface

BASE_EDD: Baseline earliest-due-date policy

BASE_FIFO: Baseline first-in, first-out policy

CNC: Computer numerical control

CPU: Central processing unit

CRP: Capacity Requirements Planning

DDR5: Double data rate 5

DSR: Design Science Research

DSS: Decision Support System

E2: Stakeholder-weighted multi-criteria evaluation model

EDD: Earliest Due Date

ERP: Enterprise Resource Planning

ETO: Engineer-to-Order

FIFO: First-In, First-Out

FP16: 16-bit Floating-Point Format

GenAI: Generative Artificial Intelligence

GPU: Graphics processing unit

HITL: Human in the Loop

IIoT: Industrial Internet of Things

IRR: Iranian rial

JSON: JavaScript Object Notation

KPI: Key Performance Indicator

LLM: Large Language Model

MRP: Material Requirements Planning

MTO: Make-to-Order

NCR: Non-Conformity Report

OTD: On-Time Delivery

OTDR: Optical Time-Domain Reflectometer

QC: Quality Control

RAG: Retrieval-Augmented Generation

RAM: Random-access memory

REST: Representational State Transfer

S&OP: Sales and operations planning

SOP: Standard Operating Procedure

STS: Sociotechnical Systems

VRAM: Video random-access memory


DOI: 10.22067/jstinp.2026.98779.1206

XAI: Explainable Artificial Intelligence

## References

An, Y., Chen, X., Gao, K., Zhang, L., Li, Y. and Zhao, Z., 2023. Integrated optimisation of real-time order acceptance and flexible job-shop rescheduling with multi-level imperfect maintenance constraints. Swarm and Evolutionary Computation, 77, p.101243. https://doi.org/10.1016/j.swevo.2023.101243.

Augustsson, H., Churruca, K. and Braithwaite, J., 2020. Change and improvement 50 years in the making: a scoping review of the use of soft systems methodology in healthcare. BMC Health Services Research, 20(1), p.1063. https://doi.org/10.1186/s12913-020-05929-5.

Bhalla, S., Alfnes, E., Hvolby, H.H. and Oluyisola, O., 2023. Sales and operations planning for delivery date setting in engineer-to-order manufacturing: a research synthesis and framework. International Journal of Production Research, 61(21), pp.7302-7332. https://doi.org/10.1080/00207543.2022.2148010.

Cortiñas-Lorenzo, K., Cai, W. and Doherty, G., 2025. Designing, implementing, and evaluating AI explanations: a scoping review of explainable AI frameworks. ACM Transactions on Computer-Human Interaction, 32(6), pp.1-79. https://doi.org/10.1145/3769678.

Françozo, R., Paucar-Caceres, A. and Belderrain, M.C.N., 2022. Combining value-focused thinking and soft systems methodology: a systemic framework to structure the planning process at a special educational needs school in Brazil. Journal of the Operational Research Society, 73(5), pp.994-1013. https://doi.org/10.1080/01605682.2021.1880298.

Garcia, C.I., DiBattista, M.A., Letelier, T.A., Halloran, H.D. and Camelio, J.A., 2024. Framework for LLM applications in manufacturing. Manufacturing Letters, 41, pp.253-263. https://doi.org/10.1016/j.mfglet.2024.09.030.

Ghobakhloo, M., Fathi, M., Iranmanesh, M., Vilkas, M., Grybauskas, A. and Amran, A., 2024. Generative artificial intelligence in manufacturing: opportunities for actualizing Industry 5.0 sustainability goals. Journal of Manufacturing Technology Management, 35(9), pp.94-121. https://doi.org/10.1108/JMTM-12-2023-0530.

Goto, Y. and Miura, H., 2022. Using the soft systems methodology to link healthcare and long-term care delivery systems: a case study of community policy coordinator activities in Japan. International Journal of Environmental Research and Public Health, 19(14), p.8462. https://doi.org/10.3390/ijerph19148462.

Hörbe, R. and Erol, S., 2024. Artificial intelligence in planning and control tasks: a study of potential use cases and perceived challenges in Austrian make-to-order companies. Procedia CIRP, 130, pp.232-237. https://doi.org/10.1016/j.procir.2024.10.081.

Jackson, I., Ivanov, D., Dolgui, A. and Namdar, J., 2024. Generative artificial intelligence in supply chain and operations management: a capability-based framework for analysis and implementation. International Journal of Production Research, 62(17), pp.6120-6145. https://doi.org/10.1080/00207543.2024.2309309.

Jeon, J., Sim, Y., Lee, H., Han, C., Yun, D., Kim, E., Nagendra, S.L., Jun, M.B.G., Kim, Y., Lee, S.W. and Lee, J., 2025. ChatCNC: conversational machine monitoring via large language model and real-time data retrieval augmented generation. Journal of Manufacturing Systems, 79, pp.504-514. https://doi.org/10.1016/j.jmsy.2025.01.018.

Kapulin, D.V. and Russkikh, P.A., 2020. Analysis and improvement of production planning within small-batch make-to-order production. Journal of Physics: Conference Series, 1515(2), p.022072. https://doi.org/10.1088/1742-6596/1515/2/022072.

Kusiak, A., 2025. Generative artificial intelligence in smart manufacturing. Journal of Intelligent Manufacturing, 36(1), pp.1-3. https://doi.org/10.1007/s10845-024-02480-6.

Leng, J., Zheng, K., Li, R., Chen, C., Wang, B., Liu, Q., Chen, X. and Shen, W., 2026. AIGC-empowered smart manufacturing: prospects and challenges. Robotics and Computer-Integrated Manufacturing, 97, p.103076. https://doi.org/10.1016/j.rcim.2025.103076.

Ma, G., da Silva, E.R. and Møller, C., 2024. Digital twin-based demand-driven production planning and scheduling system: a new conceptual framework. Procedia CIRP, 130, pp.705-710. https://doi.org/10.1016/j.procir.2024.10.152.

Mundt, C. and Lödding, H., 2025. Coping with the uncertainties of make-to-order production: a new approach for determining reliable delivery times with the throughput diagram. Production Planning & Control, 36(8), pp.1110-1136. https://doi.org/10.1080/09537287.2024.2344066.

Nikiforidis, K., Kyrtsoglou, A., Vafeiadis, T., Kotsiopoulos, T., Nizamis, A., Ioannidis, D., Votis, K., Tzovaras, D. and Sarigiannidis, P., 2025. Enhancing transparency and trust in AI-powered manufacturing: a survey of explainable AI (XAI) applications in smart manufacturing in the era of industry 4.0/5.0. ICT Express, 11(1), pp.135-148. https://doi.org/10.1016/j.icte.2024.12.001.

Sit, S.K.H. and Lee, C.K.M., 2023. Design of a digital twin in low-volume, high-mix job allocation and scheduling for achieving mass personalization. Systems, 11(9), p.454. https://doi.org/10.3390/systems11090454.

Tang, P.M., Leong, K.K.H., Shaik, N. and Lau, H.C., 2024. Automated conversion of static to dynamic scheduler via natural language. arXiv preprint arXiv:2405.06697. https://doi.org/10.48550/arXiv.2405.06697.

Wattuhewa, S. and Pheerathan, S., 2025. Reflective applications of soft systems methodology (SSM) across contexts. Systemic Practice and Action Research, 38(4), p.31. https://doi.org/10.1007/s11213-025-09748-9.